Machine Learning 1

Lecture 13.1 - Combining Models
Bayesian Averaging vs Model Combination

Erik Bekkers

(Bishop 14.0, 14.1)

Slide credits: Patrick Forré and Rianne van den Berg

Regression with GP's

‣ Combining models: (Bishop 4.1-4.4)

  ‣ **Bayesian model averaging vs. model combination methods**

  ‣ Committees:

    ‣ Bootstrap aggregation

    ‣ Random subspace methods

    ‣ Boosting

  ‣ Decision trees

  ‣ Random forests

Regression with GP's

‣ So far we have considered many different models for classification and regression

‣ It is often the case that overall performance can be improved by combining multiple models together in some way.

‣ Regression example: train L different models and make predictions using the average of the predictions made by each model.

‣ Methods that are combined like this are called **committees**

Model combination

‣ Committee example: boosting

‣ Train multiple models in sequence

‣ Error function used to train a particular model depends on performance of previous models

‣ Boosting algorithms can lead to substantial improvements over individual models!

Alternative: ==model selection==
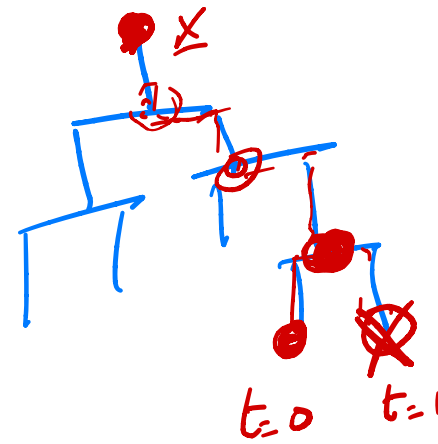
‣ For each prediction, select one model to make a prediction.

‣ ==The choice of the model that is selected is a function of the input variables==

‣ Example 1: Decision trees! Selection process is a sequence of binary selections corresponding to the traversal of a tree structure

‣ Example 2: Mixtures of experts. Soft selection of models for predictions

one of the experts

$$p(t \mid \mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}) p(t \mid \mathbf{x}, k)$$

mixing coefficients which depends on $\underline{x}$

$t=0 \quad t=1$

Bayesian model averaging

VS

Model combination methods

Bayesian model averaging vs. combining models

‣ Let's make sure we understand the difference between Bayesian model averaging and model combination methods

‣ We have already seen a model combination method for density estimation: Gaussian mixture models!

‣ Several Gaussians are combined probabilistically to produce the density $p(\mathbf{x})$

$$p(x) = \sum_{z} p(x \mid z) p(z)$$

‣ A binary latent variable $\mathbf{z}$ that indicates which component of the mixture is responsible for generating the datapoint $\mathbf{x}$.

‣ The model specifies $\quad p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z})$

with $\quad p(\mathbf{x} \mid z_k = 1) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

$$\mathbf{z} \in \mathbb{R}^K \quad, \quad \underline{\mathbf{z}} = (0, 0, 1, \cdots)^T \quad \text{(one-hot enc.)}$$
$$z_k$$

Combining models: GMM

- A Gaussian Mixture model specifies $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$

$$\text{with } p(\mathbf{x} \mid z_k = 1) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Then the density over observed $\mathbf{x}$ is obtained by

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- For i.i.d data $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1, \ldots, \mathbf{x}_N \end{bmatrix}^T$ we have

$$p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x}_n) = \prod_{n=1}^{N} \sum_{\mathbf{z}_n} p(\mathbf{x}_n, \mathbf{z}_n)$$

- So each observed datapoint $\mathbf{x}_n$ has its own latent variable $\mathbf{z}_n$!

Bayesian model averaging

‣ Suppose we have different models, indexed by $h = 1, \ldots, H$

‣ We also have prior probabilities $p(h)$

‣ Marginal distribution over the dataset is

$$p(\mathbf{X}) = \sum_{h=1}^{H} p(\mathbf{X}\,|\,h)p(h)$$

‣ Interpretation: one model is responsible for generating the entire dataset

‣ $p(h)$ simply reflects our uncertainty which model is the correct model

‣ If dataset size increases, uncertainty is reduced: $p(h\,|\,\mathbf{X})$ becomes increasingly focused on one model

Contrast with model combination methods

‣ In Bayesian model averaging: the entire dataset is generated by a single model, we are just unsure which one it was!

‣ When we combine multiple models different data points can potentially be generated by different components/models!

‣ Example in GMM:

   ‣ Take two datapoints $\mathbf{x}$ and $\mathbf{x}'$

   ‣ They can be generated from different $\mathbf{z}$ and $\mathbf{z}'$

   ‣ So they come from different model components!